Sociolinguistic variation in morphological productivity in eighteenth-century English

Säily, Tanja

2016-05

Säily , T 2016 , ' Sociolinguistic variation in morphological productivity in eighteenth-century English ' , Corpus Linguistics and Linguistic Theory , vol. 12 , no. 1 , pp. 129-151 . https://doi.org/10.1515/cllt-2015-0

http://hdl.handle.net/10138/223823
https://doi.org/10.1515/cllt-2015-0064

publishedVersion

Tanja Säily

Sociolinguistic variation in morphological productivity in eighteenth-century English

DOI 10.1515/cllt-2015-0064

Abstract: This paper presents ongoing work on Säily and Suomela's (2009) method of comparing type frequencies across subcorpora. The method is here used to study variation in the productivity of the suffixes *-ness* and *-ity* in the eighteenth-century sections of the *Corpora of Early English Correspondence* and of the *Old Bailey Corpus* (OBC). Unlike the OBC, the eighteenth-century section of the letter corpora differs from previously studied materials in that there is no significant gender difference in the productivity of *-ity*. The study raises methodological issues involving periodization, multiple hypothesis testing, and the need for an interactive tool. Several improvements have been implemented in a new version of our software.

Keywords: historical sociolinguistics, gender variation, Late Modern English, methodology, morphological productivity, word-formation, nominal suffixes, type frequency

1 Introduction

This paper is concerned with applying measures of morphological productivity to sociolinguistic variation and change in the long diachrony. Several measures of morphological productivity have been proposed by Baayen (e.g. 1993), but they are all dependent on the size of the corpus, which makes it difficult to compare measures obtained from sociolinguistically defined subcorpora of different sizes. Säily and Suomela (2009) suggest an assumption-free, highly visual solution based on type accumulation curves and the statistical technique of permutation testing. The present paper proposes several improvements to this method.

So far, the method has been applied to the study of gender variation in seventeenth-century and present-day materials (Säily and Suomela 2009; Säily 2011). Interestingly, it was seen that in both Early Modern and Present-day English writing, women used *-ity* significantly less productively than men, while there were no significant differences in the use of *-ness*. This could imply a gendered

Tanja Säily, Department of Modern Languages, University of Helsinki, Finland.
E-mail: tanja.saily@helsinki.fi

discourse style remaining stable throughout the centuries (see Nevalainen 2002: 191–194). To test the hypothesis of a stable gendered style, the investigation of the productivity of *-ity* and *-ness* is here extended into the less studied territory of eighteenth-century English. The method will also be used to study the correlation of social rank and morphological productivity.

The remainder of the paper is organized as follows. Section 2 discusses previous work on measuring morphological productivity and describes the method as it is used in this paper. Section 3 provides a brief overview of the social situation in eighteenth-century England and introduces the corpora used in the study. Section 4 presents the results of the study, while Section 5 discusses both the results and some methodological issues raised by them. Finally, Section 6 concludes the paper with remarks on the results, method, and future work.

2 Method

2.1 Previous work on measuring morphological productivity

The question of how to measure morphological productivity, or indeed how morphological productivity should be defined, has been the subject of scholarly debate for the past decades.

Baayen (1992, 1993) advocates a psycholinguistic approach, taking as a starting point the "morphological race" model of what happens in a language user's mind as he or she processes a complex word. The idea is that the user has a mental lexicon which contains both single morphemes and complex words. A complex word such as *kindness* can be either retrieved from the mental lexicon as a whole or parsed into its component morphemes, i.e. the base, *kind*, and the derivational affix, *-ness*. Which route is faster depends in part on the frequency of the complex word: a high-frequency word is more likely to be retrieved as a whole, whereas a low-frequency word is less active in the user's memory and thus more likely to be parsed. Parsing, on the other hand, maintains the activation level of the affix, facilitating production as well as perception. Therefore, the productivity of an affix can be estimated by examining the frequency spectrum of the complex words that contain it – a high proportion of low-frequency words (which need to be parsed) implies high productivity.

This has led Baayen to develop two productivity measures based on hapax legomena or hapaxes, which are words having the extremely low frequency of 1 in a given corpus. Expanding productivity, P^*, is defined as the number of hapaxes containing the affix in question divided by the overall number of

hapaxes in the corpus, $P^* = n_1/h$. Potential productivity, P, is defined as the number of hapaxes containing the affix in question divided by the number of all word tokens containing the affix in question, $P = n_1/N$. According to Baayen (2009: 902), P^* assesses "the contribution of [a] morphological category... to the growth rate of the total vocabulary", whereas P "estimates the growth rate of the vocabulary of the morphological category itself". The third facet of productivity involves the current size of the morphological category. Baayen (2009: 901–902) calls this realized productivity, V, defining it as the number of different words containing the affix in question, or type frequency.

Hay (2001) proposes a measure of morphological productivity that takes into account the frequency of the base as well as the affixed form. However, measures like this are not well suited for corpora that have not been lemmatized or tagged, such as historical corpora containing a great deal of spelling variation. Another issue with many historical corpora is their relatively small size, which prevents the use of productivity measures based on hapax legomena (Säily and Suomela 2009; Säily 2011). Moreover, both hapax- and type-based measures become problematic when we wish to compare subcorpora of different sizes, such as different time periods or social categories. This issue is discussed further in the next section.

2.2 Comparing type frequencies

A conventional way to study variation and change in productivity is to compare type or hapax frequencies across subcorpora (e.g. Dalton-Puffer 1996: 106). The problem with comparing type and hapax frequencies of any linguistic item is that these measures depend on the size of the (sub)corpus in a non-linear manner (Baayen 1992: 113; Säily 2011: 127), which means that standard techniques such as normalization are not applicable. In short, normalization assumes that the rate at which we observe new types as we progress through a corpus remains constant, whereas in reality we will encounter new types more often at the beginning of the corpus, and the rate will decrease as the size of the corpus increases (see Säily 2011: 124). Säily and Suomela (2009) present a simple method for comparing the type and hapax frequencies of an affix (or, to be more precise, the number of word types and hapaxes containing the affix) across subcorpora of any size. Unlike models based on extrapolation (see Baayen 2001; Evert and Baroni 2005), this method does not make simplifying assumptions like "words occur randomly in texts". Furthermore, it is highly visual (for the benefits of corpus visualization, see Siirtola et al. 2011) and provides a built-in measure of statistical significance.

The idea behind the method is to divide the corpus into samples large enough to preserve discourse structure (e.g. samples the size of individual texts). For each sample, the user needs to note down its size (in running words or affix tokens) and the type or hapax frequency of the affix under analysis. The samples are then picked in a random order by a computer program (Suomela 2007) to construct a type accumulation curve for the affix in the corpus. The procedure is repeated a large number of times: in this study, a million times. The million random permutations of the corpus provide upper and lower bounds for comparing the type frequencies of the affix in actual subcorpora. As an example, Figure 1 plots the type frequencies of the suffix *-ity* in gender-based subcorpora against random accumulation curves for the entire corpus. The figure shows that more than 99.9% of the randomly composed subcorpora that are the same size as the female subcorpus have a greater type frequency of *-ity* than it, making the type frequency of the female subcorpus significantly low at $p < 0.001$.

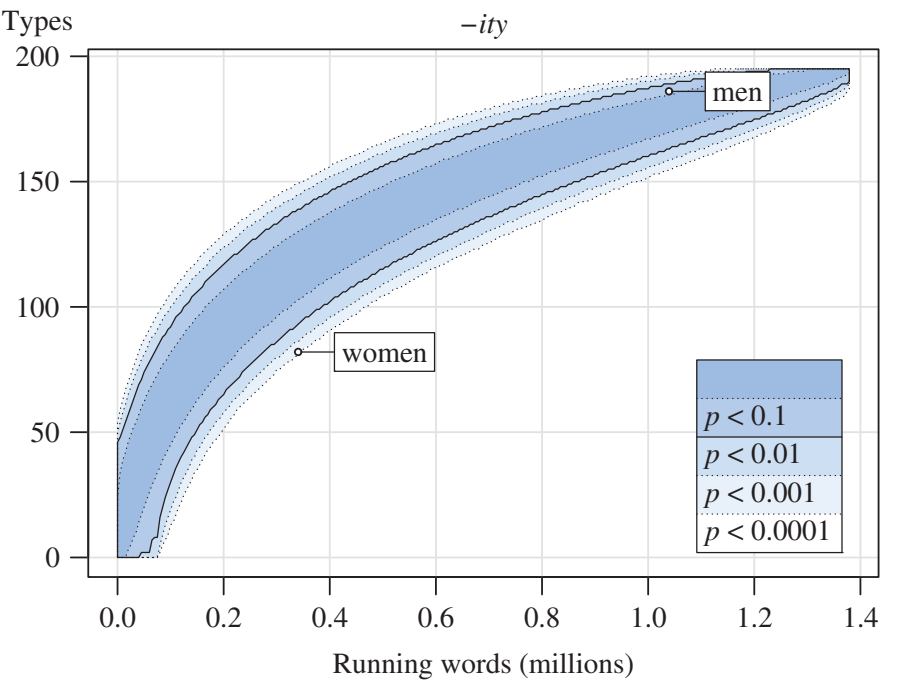

Figure 1: Bounds for 1,000,000 type accumulation curves, with gender-based subcorpora plotted on the curves, for the suffix *-ity* in the seventeenth-century part of the *Corpus of Early English Correspondence*. Based on Säily and Suomela (2009: Figure 5).

2.3 Diachronic periodization

As has been discussed by Gries and Hilpert (2008), there are many ways to divide a historical corpus into time periods. For the method described above, one way is to create the type accumulation curves for the entire period under analysis, plotting individual subperiods on the curves as desired. Sociolinguistic studies often use generation-based, twenty-year subperiods, although lack of data sometimes requires doubling these up into forty-year periods (e.g. Nevalainen and Raumolin-Brunberg 2003). These sociolinguistically motivated twenty-year subperiods are also used in the present study.

A problem with having one set of type accumulation curves for the entire corpus is that tracking change over time is difficult, as we can only get significant results for subperiods that have a very low or high type frequency in comparison with the corpus as a whole. If there is an increase in the productivity of a suffix over time, it will show up indirectly either as a significantly low type frequency for the first subperiod or as a significantly high type frequency for the last subperiod (or possibly both), and we cannot say much about the intervening periods. Therefore, the present study also creates type accumulation curves for time periods shorter than the 120 years covered by the entire corpus, namely forty- and eighty-year periods. This is done using a sliding window, with the starting points twenty years apart for the forty-year curves and forty years apart for the eighty-year curves, plotting the appropriate 20-year subperiods on each set of curves. Social categories, too, are plotted on these curves, so that any differences observed across the categories in the entire corpus can be tracked more closely.

Another way to observe productivity changes in social categories is to construct the curves for one category only and to plot subperiods on the curves. In the present study, this is done for men, women, and a few social ranks. Unfortunately, there is often too little data for significant results to emerge within either social categories or shorter time periods.

3 Material

3.1 Background: The eighteenth century

To test the hypothesis of a stable gendered style, we need to analyze data from a time period between those already studied, the seventeenth and late twentieth centuries. The corpora analyzed should preferably also be comparable with

those used in the previous studies. This leads us rather naturally to the eighteenth century, for which an extension to the original *Corpus of Early English Correspondence* (CEEC; ca.1410–1681) is available. As the compilation principles across the seventeenth- and eighteenth-century sections of the corpus are the same, the results should be easy to compare. A further speech-related genre is provided by the *Old Bailey Corpus* (OBC), which contains trial proceedings. Before examining the corpora in more detail, let us survey the setting in which the material was produced, i.e. eighteenth-century England.

How did the social situation in eighteenth-century England differ from that in the seventeenth century? According to Hay and Rogers (1997: 18–24), the crucial division in society was still between gentry and non-gentry. The line had become more blurred than before, however, so that landownership or even freedom from manual labor was no longer essential, and wealthy merchants and sons of great manufacturers could be called gentlemen. Fitzmaurice (2012) notes that the number of tradesmen and manufacturers increased dramatically during the period.

Education was more widely available but still stratified, universities being reserved for men of the "better sort" (Cannadine 2000 [1998]: 47–48). Women's education, too, was somewhat improved, so that most women of the "better sort" were literate, and some had received a high-level education at home, although this was not necessarily encouraged by society at large (Tieken-Boon van Ostade 2010). Notably, this period saw the rise of the group of educated and intellectual women known as the Bluestockings (Myers 1990; Pohl and Schellenberg 2003).

3.2 *The Corpus of Early English Correspondence* (CEEC) and its *Extension* (CEECE)

As a follow-up to Säily and Suomela's (2009) seventeenth-century study, which used the original CEEC ending in 1681, the present study examines English correspondence in the long eighteenth century, 1680–1800. Hence, the tail end of the CEEC (1680–1681) is here combined with the *Corpus of Early English Correspondence Extension* (CEECE), which contains letters chiefly from 1680–1800. The data set consists of 4,946 letters written by 313 people, or ca. 2.2 million words.

Based on edited letter collections, the CEEC family of corpora was designed for the purposes of historical sociolinguistics. Therefore, the sampling unit was the individual letter writer, and an effort was made to create a balanced corpus in terms of gender, social status, and time period. As there was less material available from women, the lower ranks, and the earlier periods, this goal was not achieved completely, but the CEEC and CEECE nevertheless remain among the best resources available for this type of study. Because the genre – personal letters – is in some

respects "speech-like" (Culpeper and Kytö 2010: 17), it is a promising source for tracking language change. Furthermore, the rich sociolinguistic metadata associated with the corpora facilitates the analysis of various social categories; among them are gender and social rank, which are the focus of this study.

In the CEEC family of corpora, the basic division of people into social ranks has been kept the same for the entire period of ca.1410–1800: royalty, nobility, gentry, clergy, professionals, merchants, and other non-gentry (Raumolin-Brunberg and Nevalainen 2007). At the same time, it has been taken into account that by the eighteenth century, the lines between social ranks had become more blurred (see Section 3.1); accordingly, some artists and manufacturers have been counted as being members of the rank of professionals rather than of the lowest, "other" category. For the most part, unmarried women inherit their social rank from their fathers and married women from their husbands, which is a common practice even in present-day sociolinguistic studies.

3.3 *Old Bailey Corpus* (OBC)

As a complement to the correspondence corpora, a different genre from the eighteenth century is examined, namely that of trial proceedings. Developed at the University of Giessen, the OBC is based on the proceedings of the Old Bailey, London's central criminal court, which were published from 1674 to 1913. Version 0.4 of the corpus includes ca. 4.1 million words of spoken material from the eighteenth century, 1730–1800. The gender of the speakers is known for ca. 3.2 million words, and social rank in addition to gender for ca. 0.5 million words.

The advantage to the corpus is that it is arguably the closest we can get to the spoken language of the time, as the texts are based on actual speech events. Furthermore, it provides access to the language of the lower classes (Culpeper and Kytö 2010: 16–17). However, as noted by Huber (2007: Section 5), in a corpus of trial proceedings "what looks like language variation and change may in fact be due to the influence of scribes and printers". This needs to be taken into account when assessing the results.

4 Results

4.1 CEEC and CEECE

As can be seen in Figure 2, the most significant result for *-ity* is its low productivity (in terms of type frequency as a function of the number of running

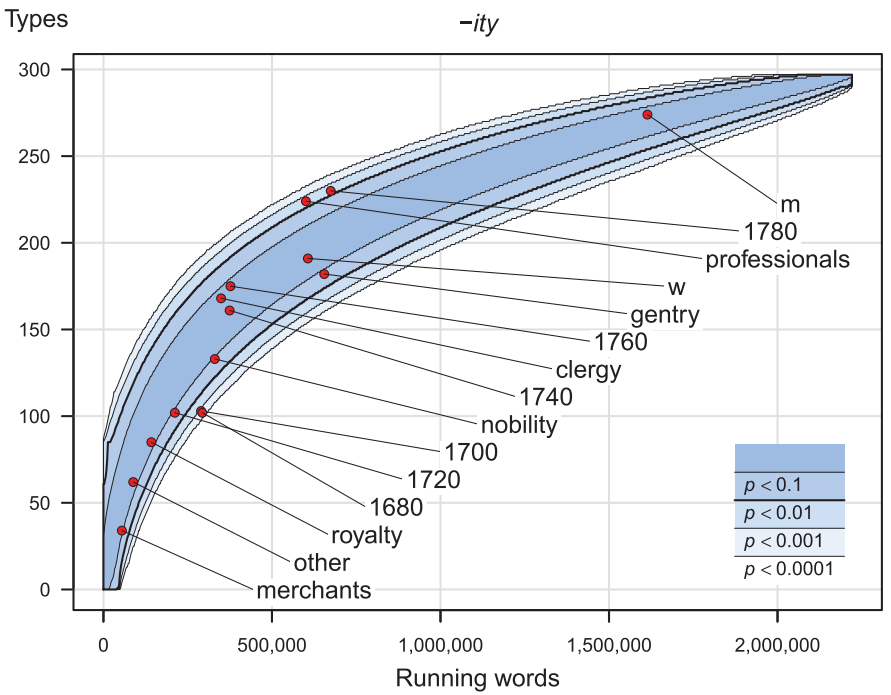

Figure 2: Sociolinguistic variation and change in the productivity of *-ity* in the CEEC + CEECE, 1680–1800 (type frequency as a function of the number of running words).

words) in the earliest two subperiods, 1680–1699 and 1700–1719. Moreover, its productivity is significantly high in the last subperiod, 1780–1800. The subperiods in between are not significantly different from the corpus as a whole, but the productivity of 1720–1739 is lower (compared to all randomly composed subcorpora of the same size) than that of 1740–1759, whose productivity in turn is lower than that of 1760–1799.[1] All in all, these results clearly indicate an increase in the productivity of *-ity* over time, which was also the case in the seventeenth-century part of the CEEC (Säily and Suomela 2009).

1 In the accumulation curves constructed for shorter time periods (see Section 2.3 above), the amount of data is for the most part too small for significant results to emerge, but the change is observable in the eighty-year stretch of 1680–1759 as a significantly high productivity of *-ity* in the final subperiod, 1740–1759.

Also plotted on Figure 2 are subcorpora based on gender and social rank. Our hypothesis of a stable gendered style would have predicted a significantly low productivity of *-ity* with women in eighteenth-century correspondence (maintaining the level of significantly low productivity observed in the seventeenth century); this is not borne out by the facts, as the productivity of *-ity* is not significantly low with women in 1680–1800. What does emerge as significant is its high productivity in the social rank of professionals, such as government officials, doctors, lawyers, and authors (see (1), (2)).[2] A possible explanation for this is that their topics of writing required abstract nouns, and the use of "hard" words (as they were known at the time) with the Latinate/Romance suffix *-ity* gave them the opportunity to show off their prestigious classical education. This would then seem to be a question of style and the cultural norms of polite society, to which professionals aspired to belong (Nevalainen and Tissari 2010: 141, 146–147).

(1) *Most of my **fraternity** would as soon shorten the noses of their children because they were said to be too long, as thus dock their compositions in compliance with the opinion of others. I beg that when my life shall be written hereafter my Authorship's **ductility** of temper may not be forgotten.*

(COWPERW_063, *William Cowper to Walter Bagot*, 1789)

(2) *Even yet, perhaps, your interest & influence (could you feel that **security** & reliance in his given honour that I do, so as to act for him warmly) might draw him from **obscurity** & penury, to be of service to his Country, &, with his admirable professional talents, of use & honour to his Family.*

(BURNEYF_068, *Fanny Burney to her father*, 1800)

When looking at the type frequency of *-ity* as a function of the number of suffix tokens rather than the number of running words, the results are similar but less significant (Figure 3). This has also been the case in previous studies (Säily and Suomela 2009; Säily 2011).

Let us now turn to the suffix *-ness*. As can be seen from Figure 4, there are no significant differences in terms of gender, social rank, or time period when productivity is measured in terms of type frequency as a function of the number of running words. The situation is similar to that in the seventeenth

2 There is not enough data to compare male and female professionals or to detect change over time within the ranks.

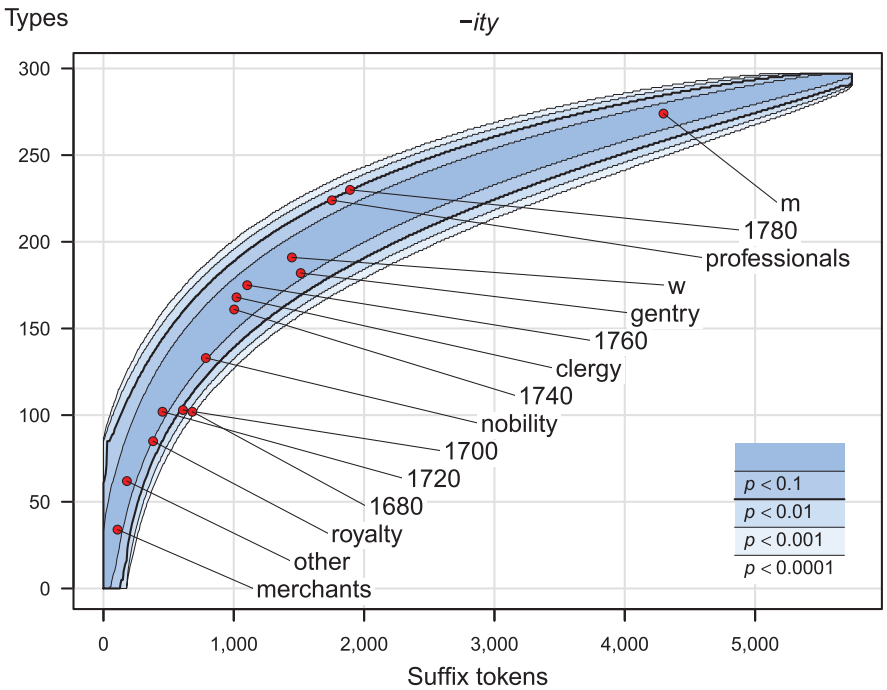

Figure 3: Sociolinguistic variation and change in the productivity of *-ity* in the CEEC + CEECE, 1680–1800 (type frequency as a function of the number of suffix tokens).

and twentieth centuries and would thus seem to support the hypothesis of stability.

Surprisingly, however, when the *x*-axis is switched to show the number of suffix tokens, two differences become significant: the productivity of *-ness* is significantly low with royalty and significantly high with clergy (Figure 5).[3]

A low productivity of *-ness* in terms of type frequency as a function of the number of suffix tokens means that the same words in *-ness* are repeated over and over. With royalty, this is probably due to the fact that many of the letters were written by junior members of the royal family to senior members, flattering or thanking the recipients in a formulaic way (3) and using specific closing formulae (4) to show respect and create goodwill. Some of the letters

3 Again, there is not enough data to compare male and female members of these ranks or to detect change over time within the ranks.

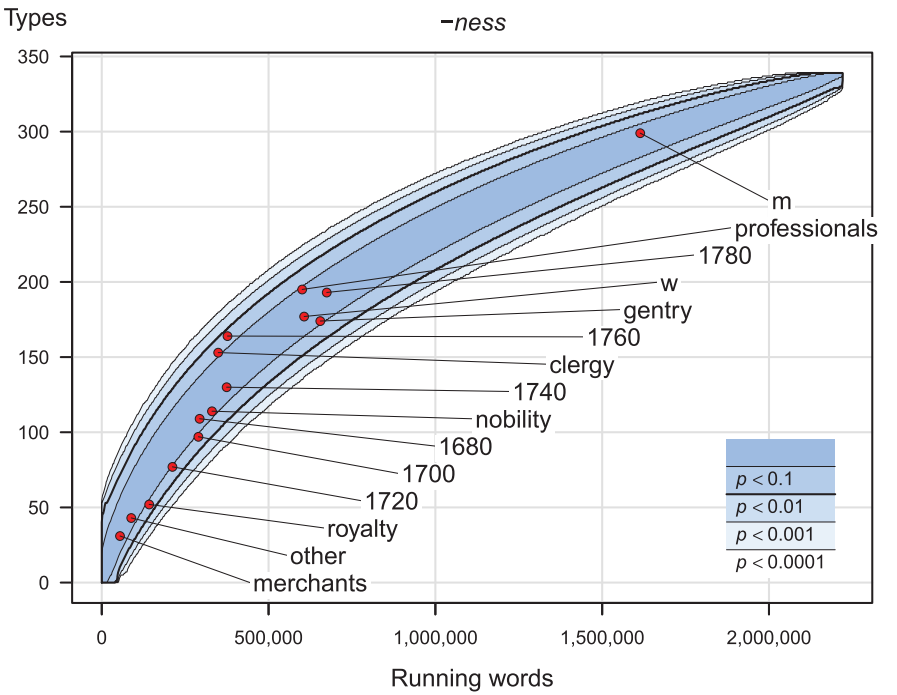

Figure 4: Sociolinguistic variation and change in the productivity of *-ness* in the CEEC + CEECE, 1680–1800 (type frequency as a function of the number of running words).

may have served no other purpose than an exercise in rhetoric (see Nevalainen 2009: 142).

(3) *Your Majesty's **goodness** and **kindness** towards me gives me great hopes of this fortunate event soon taking place. My petition now is...*
(GEORG3A_071, *Prince Augustus to the King*, 1795)

(4) *I must humbly beg of your Majesty to present my most respectful duty to the Queen, and if I might presume to request it of your Majesty, my most affectionate love to my sisters who, I trust, will ever join me in prayer to Heaven for your **happiness** & that of the Queen.*
(GEORG3A_010, *Prince Edward to the King*, 1785)

The high productivity of *-ness* with clergy is perhaps a more interesting issue, and may be related to the high frequency of *-ness* types in sermons, a text type

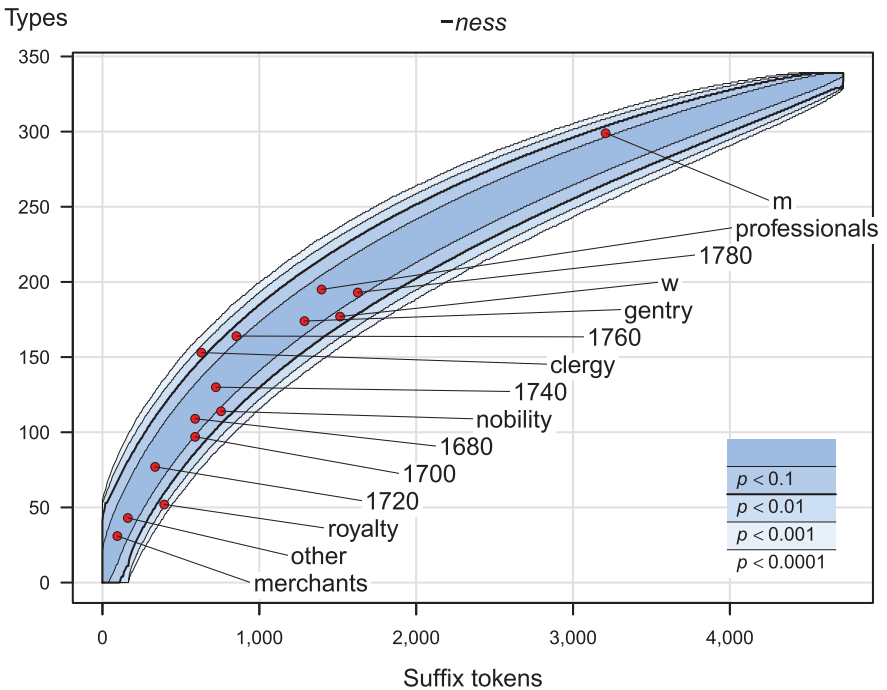

Figure 5: Sociolinguistic variation and change in the productivity of *-ness* in the CEEC + CEECE, 1680–1800 (type frequency as a function of the number of suffix tokens).

produced by clergymen. Indeed, sermons are reported by Cowie (1999: 239) as having contributed the most new *-ness* types to the ARCHER corpus (1650–1990), which comprises ten different registers. According to Cowie (1999: 242), late seventeenth-century sermons used *-ness* to ensure that everyone could understand the message. While audience comprehension would not have been an issue to clergymen writing letters to people in their social circle, it is possible that the habit of using *-ness* still carried over from their sermons to their correspondence.[4] In addition, they would sometimes quote or paraphrase religious texts containing *-ness* in their letters, as in (5).

4 This reasoning does not apply to women belonging to the social rank of clergy, as (apart from nuns) they only belonged to it by virtue of being daughters or wives of clergymen, and they did not as a rule write sermons; however, there is so little data from them that it does not have much of an effect on the results.

(5) *... they also sprinkl'd her & y^e Bed & room with their holy water & fel a sweeping y^e room with besoms, as hard as they could, to sweep all her sins away; ô that ever there should bee such* ***darkness*** *in the midst so much light!*
(HENRY_014, *Philip Henry to his son*, 1687)

The letters of the clergy were not all about religious discourse, however. A few individuals used *-ness* in quite a creative or even playful manner, as in (6).

(6) *He wrote me, on the occasion, 2 or 3 pages of most manly* ***inside-outness*** *& impartiality, such as hardly ever came, I believe, from any man but himself.*
(TWINING_044, *Thomas Twining to his brother*, 1788)

It could be argued that individual outliers from whom a large number of letters are present in the corpus may have skewed the results. However, this is made less likely by the fact that the samples used to construct the accumulation curves are not single letters, in which case an outlier with a large number of letters could easily skew the results, but consist of an individual's letters from a twenty-year period. Thus, an individual author only contributes a few samples at most, and cannot have a severely disproportionate effect.

4.2 OBC

Let us first consider *-ity* again. The absence of a significant gender difference in the use of *-ity* in the eighteenth-century part of the correspondence corpora seems to disprove the hypothesis of a stable gendered style. But could this be an artifact of the material? To explore this question, let us turn to a different corpus of eighteenth-century English, the *Old Bailey Corpus*.

As can be seen in Figure 6, the OBC seems to fall in line with results from seventeenth-century and present-day data in that the productivity of *-ity* is significantly low with women. There are some complications with the material, however. Firstly, there is a substantial amount of material from persons of unknown gender (marked with a question mark in the figure), who use *-ity* significantly productively. Still, even if they are left out, the result for women remains highly significant. Secondly, judges and other officials could only be men, so the gender difference should be adjusted for the difference between court officials and laypeople. If the data is narrowed down to victims, witnesses, and defendants only, the result for women is still significant, but less so than in the full data set (Figure 7).

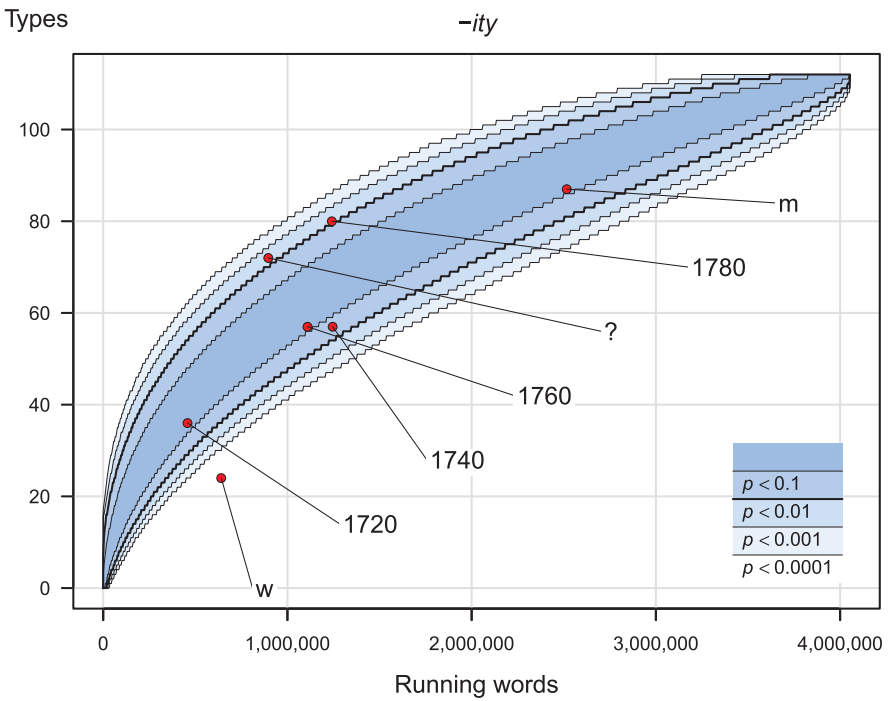

Figure 6: Sociolinguistic variation and change in the productivity of *-ity* in the eighteenth-century part of the OBC (type frequency as a function of the number of running words).

While men use both rare (7) and common words in *-ity*, women's usage is more limited to the most common types (8). The word *similarity* (7) only occurs three times in the entire corpus, whereas *opportunity* (8) is the second most frequent *-ity* word in the corpus at 346 tokens.

(7) *There was some **similarity** but not such as to deceive a clerk, unless it was overlooked; it is not well done.*
(t17860719-31, *John Wilkinson (bank teller, victim)*, 1786)

(8) *She was then on the threshold of the door; then she had, as I thought, an **opportunity** to drop them: I kept hold of her arm and cloak and pulled her, and said she should go into the back parlour: She did not seem unwilling to go.*
(t17640222-47, *Hannah Crosby (milliner, victim)*, 1764)

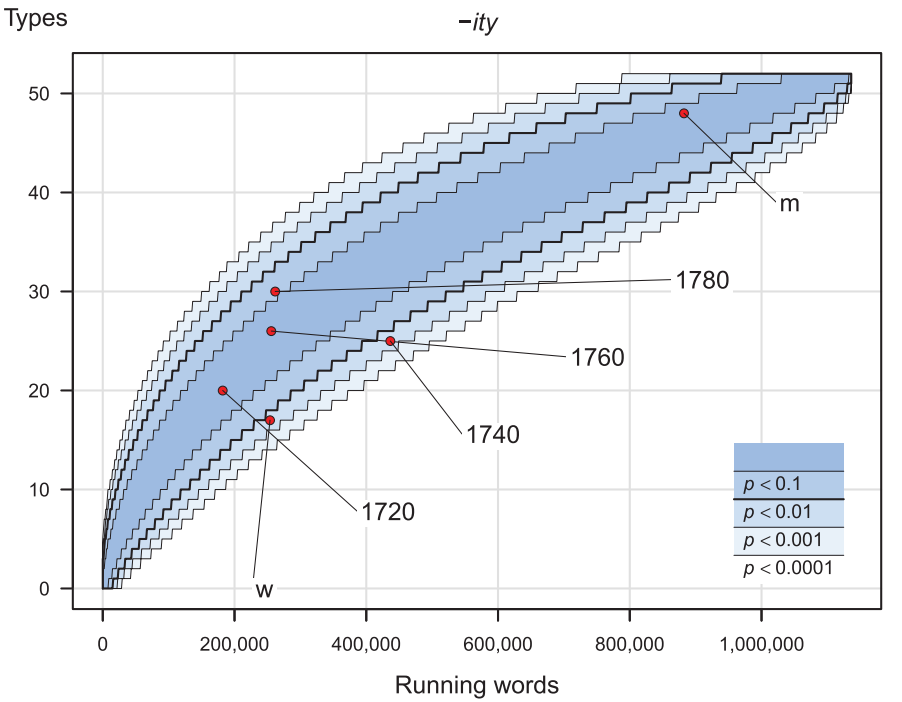

Figure 7: Sociolinguistic variation and change in the productivity of *-ity* among laypeople in the eighteenth-century part of the OBC (type frequency as a function of the number of running words).

As for *-ness*, the situation is again similar to that in the seventeenth and twentieth centuries in that there are no statistically significant gender differences (or change over time). This holds for both the full data set (Figure 8) and for laypeople only. It seems that both men (9) and women (10) use *-ness* quite diversely.

(9) *For Bakers are not accountable to the Meal-men for their Sacks, and 'tis very common for them to lose a great many by the* ***Roguishness*** *of the Journeymen Bakers, and I know 'tis Strutt's Sack, for I deal with him for Flower.*
(t17400709-39, *Arthur Findon (baker?, victim)*, 1740)

(10) *Indeed I am positive to her, and am certain she is the Lady, for I never remarked any Person more,– the* ***Agreeableness*** *of the Lady made me remark her, though to be sure we have many agreeable Persons married at our House.*
(t17400709-24, *Mary Crosier (pub keeper, witness)*, 1740)

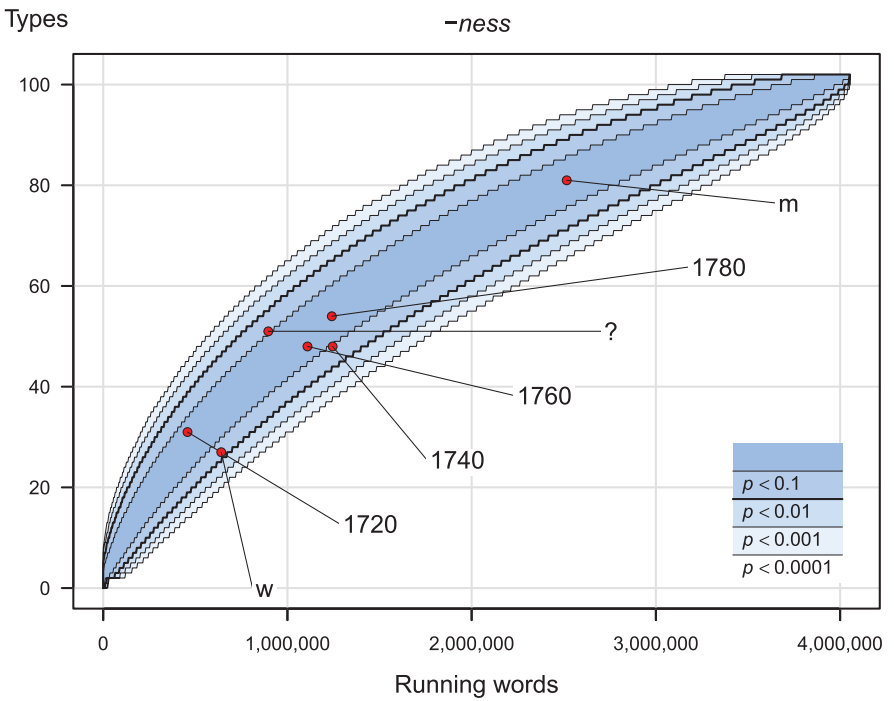

Figure 8: Sociolinguistic variation and change in the productivity of *-ness* in the eighteenth-century part of the OBC (type frequency as a function of the number of running words).

For both suffixes, when the measure of corpus size is switched from the number of running words to the number of suffix tokens, no additional significant results emerge. In the case of *-ity*, the resulting plot looks very similar to when running words are used, but the gender difference does not quite reach significance as the confidence intervals are too wide.

Thus, unlike the eighteenth-century part of the correspondence corpora, the OBC seems to behave similarly to the seventeenth-century and present-day data, supporting the hypothesis of a stable gendered style.[5]

5 There is too little data in this version of the corpus to draw conclusions on variation in the productivity of *-ness* and *-ity* in terms of social rank. The recently published version 1.0 should provide more data, enabling comparisons of this kind.

5 Discussion

5.1 Why is eighteenth-century correspondence different?

Raumolin-Brunberg and Nevalainen (2007) write that "[s]omewhat surprisingly, when compiling the CEECE we also encountered serious problems in finding lower-ranking informants from the 18th century" and that "[t]he 18th century also witnessed a new type of letter-writers, people active in literary circles". These insights into the makeup of the corpus may help to explain the lack of a gender difference in the use of *-ity*. Perhaps the educated, culturally homogeneous literati overrepresented in the corpus had developed a style of their own, a nominal style shared by both men and women.

This hypothesis is supported by Biber and Finegan's (1997) study of the ARCHER corpus, which found that eighteenth-century letters contained more elaborated (as opposed to situation-dependent) references than in the previous century and that one of the features of this style were nominalizations (see also Biber 1988: Chapter 5). The style, then, seems to be there, but was it really shared by both genders? Women's access to higher education was still more restricted than men's, and they could not participate in the unifying experience of the university. The question remains, then, whether the interaction between the genders in the private sphere was enough to create a shared style so different from what came both before and after this period.

In both the BNC and the seventeenth-century section of the CEEC, a gender difference has been observed not only in the productivity of *-ity* but also in the use of nouns in general, such that women consistently use fewer nouns and more personal pronouns than men (Rayson et al. 1997; Argamon et al. 2003; Säily et al. 2011).[6] It would be interesting to examine whether the latter difference disappears in the CEECE, which unfortunately has not yet been tagged for parts of speech. Another avenue worth pursuing would be to investigate the relationships between the senders and recipients of the letters, as letters written to family members could differ from those sent to friends and other acquaintances. The gender issue could also be clarified through a more qualitative analysis of the suffix types used in male- and female-authored letters in the seventeenth and eighteenth centuries. These tasks are left for future research.

6 For a tentative explanation of these differences and why there is no gender difference in the use of *-ness*, see Säily (2011: 130–133).

5.2 Methodological issues

In exploratory data analysis, we test for a number of possibly interesting factors. The problem with this kind of analysis is that the greater the number of observations (in this case, type counts from sociolinguistically defined subcorpora), the higher the likelihood that some of them will be significantly different by chance. This is a separate issue from testing the significance of individual observations, which is already done by our method. A simple and powerful way of taking multiple hypothesis testing into account is false discovery rate (FDR) control using the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995), which could be incorporated in our method. A preliminary application of the procedure to the present results indicates that most of them remain significant.

A further drawback of our method is that while it can control for corpus size in either running words or suffix tokens, it cannot consider both at the same time, as the *x*-axis of the curves will always be either one or the other. The current solution is to create images for both definitions of corpus size, open them in a viewer, and toggle between them for comparison. The differences might be better illustrated with a three-dimensional approach, or with a more interactive way of dealing with the two-dimensional images.

Finally, with a method like ours that is primarily visual, the volume of images created quickly becomes difficult to manage: for instance, the present study on the CEEC corpora ended up with 620 PDF images of accumulation curves, each with several subcorpora plotted on the curves. Sifting through these to find patterns of interest is cumbersome to say the least. Again, a more interactive approach is called for.

To alleviate these issues, Jukka Suomela and I have developed an improved version of the software used to produce the accumulation curves (Suomela 2007). The new version (Suomela 2014) provides actual *p*-values in addition to confidence intervals and thus a way to control FDR; furthermore, the curves are implemented as interactive SVG (Scalable Vector Graphics) images embedded on web pages with links to the other images and to the underlying data. Other features include an SQL (Structured Query Language) database for the data sets and results, and possibly a measure of effect size in a future release (see Gries 2006).

As noted by an anonymous reviewer, another way to deal with the issues of multiple hypothesis testing and the resulting high volume of images would be to abandon the exploratory approach and restrict the analysis to a few hypotheses based on previous research. Thus, the present study could have tested for gender difference alone, ignoring rank and time. However, this approach would have missed, among other things, the unexpected result that eighteenth-century letters exhibit sociolinguistic variation in the use of *-ness*, which

has been seen as the default suffix used equally by everyone to derive abstract nouns from adjectives.

In any case, previous research often suggests a number of categories worth testing: in historical sociolinguistics, the parameters of rank and time are staples of variation, so it would seem strange to omit them even if there was no particular hypothesis considering them and the phenomenon in question. Furthermore, most of the hypotheses tested in this study were variations on a theme: as an example, the gender hypothesis was tested in the corpus as a whole and within various ranks and time periods, using both measures of corpus size. This quickly adds up but is unavoidable if we wish to make sure that the result holds up in a variety of situations. For instance, had the other measure of corpus size (i.e. the number of suffix tokens) been left out on the basis of earlier studies, we would have missed the discovery that it actually matters in some cases (see Section 6 below).

6 Conclusion

The results of this study appear to lead to conflicting interpretations. On the one hand, the OBC seems to support the hypothesis of a stable gendered style in the use of *-ity*, the productivity of which is significantly low with women. On the other hand, the lack of a gender difference in the eighteenth-century section of the CEEC corpora indicates that male and female literati may have developed a shared style of letter-writing in this century. This is especially apparent in the social class of professionals, who overuse *-ity* in comparison with the corpus as a whole. Perhaps there is a general tendency for women to underuse *-ity* compared to men, but this tendency may be overridden by specific groups at specific times in specific genres. The literati of the eighteenth century in the CEEC corpora are one such group, but the literati of the late twentieth century in the BNC follow the general tendency (Säily 2011). To confirm the tendency, further corpora need to be analyzed, including data from the nineteenth century.

The method used in this paper facilitates the study of productivity variation even in small, unlemmatized historical corpora. With it, we are able to test sociolinguistic hypotheses in the long diachrony. Because hapax-based productivity measures require a large amount of data, a small corpus only allows us to study variation in the realized productivity (i.e. type frequency) of affixes and not in their growth rate; however, a diachronic corpus allows us to study real-time change in the realized productivity of affixes, which gives us a more complete picture of their productivity. Still, the scarcity of data may prevent a fine-grained analysis of variation and change within social categories.

This study has also shown that the measure of corpus size matters. When the number of suffix tokens has been employed as the measure of corpus size in previous studies, results have been similar to but less significant than when employing the number of running words (Säily and Suomela 2009; Säily 2011). This would seem to indicate that the difference between the measures lies solely in the amount of data available to the method, with more data meaning higher significance. By contrast, some of the results yielded by the present study are only significant when corpus size is measured in suffix tokens. The measure of running words in a way conflates two issues, namely, how often (in tokens) and how diversely (in types) a suffix is used, whereas the measure of suffix tokens concentrates on diversity alone.[7] Both measures may be of interest.

As the method is robust and assumption-free, it could be used as a benchmark for parametric models. Furthermore, the method could be applied to other topics besides morphological productivity and vocabulary richness. Even though the motivation for it stems from the fact that type frequencies cannot be normalized, nothing prevents the application of the method to word-frequency studies in general, with tokens rather than types on the *y*-axis. Indeed, the built-in measure of statistical significance would be highly useful as the standard corpus-linguistic tests of significance are often too simplistic, ignoring the dispersion of the words throughout the corpus (Lijffijt et al. 2012).

Possible improvements to the method include applying FDR control to the results and developing an interactive tool to facilitate comparisons. As noted in the previous section, these features have been implemented in a new version of our software (Suomela 2014), which is freely available to the community.

Acknowledgements: I am grateful to Terttu Nevalainen and Jukka Suomela for discussions and assistance, and to anonymous reviewers and the editors of this special issue for helpful feedback. I would also like to thank the participants at the ISLE 2 workshop "Current methods in English diachronic linguistics", organized by Martin Hilpert and Hubert Cuyckens, for comments on an earlier version of this paper. Thanks to Anne Gardner and Claire Cowie for their help with conquering the clergy.

7 The high productivity of *-ness* with clergy when the latter measure is employed thus means that they do not use the suffix conspicuously more often than others, but when they do use it, they use it diversely. Conversely, the low productivity of *-ness* with royalty implies that they do not use the suffix less often than others, but that the repertoire of *-ness* words they choose to use is more limited.

Funding: This research was supported in part by Langnet, the Finnish Graduate School in Language Studies.

References

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3). 321–346.

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 1993. On frequency, transparency and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 2001. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin: Mouton de Gruyter.

Benjamini, Yoav & Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1). 289–300.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253–275. Helsinki: Société Néophilologique.

Cannadine, David. 2000 [1998]. *Class in Britain*. London: Penguin Books.

Cowie, Claire. 1999. *Diachronic word-formation: A corpus-based study of derived nominalizations in the history of English*. Cambridge: University of Cambridge PhD dissertation.

Culpeper, Jonathan & Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

Dalton-Puffer, Christiane. 1996. *The French influence on Middle English morphology: A corpus-based study of derivation*. Berlin: Mouton de Gruyter.

Evert, Stefan & Marco Baroni. 2005. Testing the extrapolation quality of word frequency models. In Pernilla Danielsson & Martijn Wagenmakers (eds.), *Proceedings of Corpus Linguistics 2005*. http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx (accessed 15 October 2012).

Fitzmaurice, Susan. 2012. Social factors and language change in eighteenth-century England: The case of multiple negation. *Neuphilologische Mitteilungen* 113(3). 293–321.

Gries, Stefan Th. 2006. Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 191–202.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3(1). 59–81.

Hay, Douglas & Nicholas Rogers. 1997. *Eighteenth-century English society: Shuttles and swords*. Oxford: Oxford University Press.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6). 1041–1070.

Huber, Magnus. 2007. The Old Bailey Proceedings, 1674–1834: Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In Anneli Meurman-Solin & Arja Nurmi (eds.), *Annotating variation and change* (Studies in Variation, Contacts and Change in English 1). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/01/huber/ (accessed 9 May 2014).

Lijffijt, Jefrey, Tanja Säily & Terttu Nevalainen. 2012. CEECing the baseline: Lexical stability and significant change in a historical corpus. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen (eds.), *Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources* (Studies in Variation, Contacts and Change in English 10). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/10/lijffijt_saily_nevalainen/ (accessed 9 May 2014).

Myers, Sylvia Harcstark. 1990. *The Bluestocking circle: Women, friendship, and the life of the mind in eighteenth-century England*. Oxford: Clarendon Press.

Nevalainen, Terttu. 2002. Language and woman's place in earlier English. *Journal of English Linguistics* 30(2). 181–199.

Nevalainen, Terttu. 2009. Grasshoppers and blind beetles: Caregiver language in Early Modern English correspondence. In Arja Nurmi, Minna Nevala & Minna Palander-Collin (eds.), *The language of daily life in England (1400–1800)*, 137–164. Amsterdam: John Benjamins.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Pearson Education.

Nevalainen, Terttu & Heli Tissari. 2010. Contextualising eighteenth-century politeness: Social distinction and metaphorical levelling. In Raymond Hickey (ed.), *Eighteenth-century English: Ideology and change*, 133–158. Cambridge: Cambridge University Press.

Pohl, Nicole & Betty A. Schellenberg (eds.). 2003. *Reconsidering the Bluestockings*. San Marino, CA: Huntington Library.

Raumolin-Brunberg, Helena & Terttu Nevalainen. 2007. Historical sociolinguistics: The Corpus of Early English Correspondence. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and digitizing language corpora*, vol. 2, *Diachronic databases*, 148–171. Houndsmills: Palgrave Macmillan.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1). 133–152.

Säily, Tanja. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7(1). 119–141.

Säily, Tanja, Terttu Nevalainen & Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2). 167–188.

Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus linguistics: Refinements and reassessments*, 87–109. Amsterdam: Rodopi.

Siirtola, Harri, Terttu Nevalainen, Tanja Säily & Kari-Jouko Räihä. 2011. Visualisation of text corpora: A case study of the PCEEC. In Terttu Nevalainen & Susan M. Fitzmaurice (eds.), *How to deal with data: Problems and approaches to the investigation of the English language over time and space* (Studies in Variation, Contacts and Change in English 7).

Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/07/siirtola_et_al/ (accessed 9 May 2014).
Suomela, Jukka. 2007. *types*1: Type and hapax accumulation curves. Computer program. Zenodo. DOI: 10.5281/zenodo.9860 (accessed 9 May 2014).
Suomela, Jukka. 2014. *types*2: Type and hapax accumulation curves. Computer program. Zenodo. DOI: 10.5281/zenodo.9868 (accessed 9 May 2014).
Tieken-Boon van Ostade, Ingrid. 2010. Eighteenth-century women and their norms of correctness. In Raymond Hickey (ed.), *Eighteenth-century English: Ideology and change*, 59–72. Cambridge: Cambridge University Press.

Corpora

CEEC = *Corpus of Early English Correspondence*. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi & Minna Palander-Collin at the Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/ (accessed 15 October 2012).
CEECE = *Corpus of Early English Correspondence Extension*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily & Anni Sairio at the Department of Modern Languages, University of Helsinki.
OBC = *Old Bailey Corpus*, version 0.4. Based on Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard & Jamie McLaughlin et al., *The Old Bailey Proceedings Online, 1674–1913*. Compiled by Magnus Huber & team at the Department of English, University of Giessen. http://www.uni-giessen.de/oldbaileycorpus/ (accessed 15 October 2012).